Operations Research for Health Care 8 (2016) 42–52

Contents lists available at ScienceDirect

Operations Research for Health Care

journal homepage: www.elsevier.com/locate/orhc

Time-dependent stochastic methods for managing and scheduling Emergency Medical Services

J.L. Vile *, J.W. Gillard, P.R. Harper, V.A. Knight

School of Mathematics, Cardiff University, Cardiff, UK

ARTICLE INFO

Article history:
Received 31 October 2014
Accepted 18 July 2015
Available online 24 August 2015

Keywords:
Health care modelling
Forecasting
Priority queueing theory
Time-dependent queueing theory
Ambulance allocation
Demand and capacity models

ABSTRACT

Emergency Medical Services (EMS) are facing increasing pressures in many nations given that demands on the service are rising. This article focuses in particular on the operations of the Welsh Ambulance Service Trust (WAST), which is the only organisation that provides urgent paramedical care services on a day-to-day basis across the whole of Wales. In response to WAST's aspiration to improve the quality of care it provides, this research investigates several interrelated advanced statistical and operational research (OR) methods, culminating in a suite of decision support tools to aid WAST with capacity planning issues. The developed techniques are integrated in a master workforce capacity planning tool that may be independently operated by WAST planners. By means of incorporating methods that seek to simultaneously better predict future demands, recommend minimum staffing requirements and generate low-cost rosters, the tool ultimately provides planners with an analytical base to effectively deploy resources. Whilst the tool is primarily developed for WAST, the generic nature of the methods considered means they could equally be applied to any service subject to demand that is of an urgent nature, cannot be backlogged, is heavily time-dependent and highly variable.

1. Introduction

The Welsh Ambulance Service Trust (WAST) is the only organisation that provides urgent paramedical care services on a day-to-day basis across the whole of Wales, and as such aims to deliver high quality care wherever and whenever it is needed. Operated from 86 ambulance stations, 3 control centres and 3 regional offices, the Trust's current goal is to move away from being perceived as simply a transport service to a provider of high quality health care and scheduled transport services [1]. Facing ever increasing pressures to provide rapid responses that satisfy the targets set by the government (defined in Section 2) in the midst of a challenging two decades over which the ambulance service has seen demand levels rise threefold, WAST has been scrutinised in respect of performance issues [2,3]. Since a review of the service in 2006 however, WAST has become a much improved organisation, but still has some way to go in terms of achieving consistency across Wales and over time [4].

As WAST furthers its ambitions to provide high quality health care, it has become keen to work with partner organisations to address the issues it faces across the health service and develop new initiatives to improve its performance, resulting in a successful working relationship being established between the operational research (OR) department at Cardiff University and WAST. Upon commencement of this project, a comprehensive database was provided by the Trust comprising around 2,500,000 data records from April 2005 to December 2009, corresponding to either a submission of request for WAST assistance, the dispatch of a response vehicle, or both.

The main challenges envisioned by the Trust for the future may be classified into two distinct fields: (i) capacity planning; and (ii) location analysis. This paper addresses the primary capacity planning issue through the development of a workforce planning tool which integrates forecasting, priority queueing theory and scheduling methods into a decision support system (DSS) to optimise resource allocation in terms of capacity. The results of related investigations performed at Cardiff University to reveal improvements that could be gained from positioning resources in different locations and to model distinct identifiable parts of the ambulance cycle time are further presented in [5,6].

Personnel scheduling problems have attracted the attention of Operational Researchers for decades, who have proposed various approaches to meet potentially conflicting objectives of low operating costs and high service quality (see [7–9]). In particular, health care scheduling decisions are often highly

* Corresponding author.
E-mail address: VileJL@cf.ac.uk (J.L. Vile).

http://dx.doi.org/10.1016/j.orhc.2015.07.002

constrained problems since most services need to be assured on a continuous basis, twenty-four hours a day, seven days a week [10]. Ensuring that the right number of staff is scheduled to meet an uncertain, time-varying demand for service involves decisions about forecasting demand, acquiring capacity and deploying resources [11]. Although the integration of these four processes facilitates the creation of an optimised roster, they are often addressed separately in the literature since the methodology described allowing the decomposition of the task into several distinct parts makes the problem more tractable. For example, whilst intensive research has been conducted in the field of demand forecasting; relatively little work has been initiated in incorporating these forecasts in vehicle deployment and staffing models [12], as these models often assume that demand is known as a precursor (sometimes based on coarse ad hoc estimates [13]). Yet for these deployment schemes to be effective, it is essential that the values in the demand forecasts are accurate [14]. In order to allow the generation of a roster in a computationally efficient manner, our research addresses each task in a step by step procedure, but importantly ultimately amalgamates the processes together in a self-contained DSS for WAST.

Whilst the methods developed in this paper primarily promote efficient allocation of WAST resources, they importantly also contribute to the several fields of OR literature with novel applications and extensions of traditional forecasting methods and time-dependent priority queueing techniques. From an OR perspective, the unique linking together of the techniques in a planning tool which further captures time-dependency and two priority classes enables this research to outperform previous approaches, which have generally only considered a single class of customer [15–17] or generated staffing recommendations using approximation methods that are reliable under limited conditions [18]. The major contributions of this work are as follows:

- We incorporate forecasts of future demand generated by Singular Spectrum Analysis, SSA (a powerful nonparametric technique that appropriately deals with the stochastic nature of demand), as input to scheduling models.
- We develop efficient approximate methodologies for converting demand profiles to minimum staffing requirements in dual-class time-dependent systems, so that the proportion of urgent and routine customers subjected to excessive waits is capped.
- For situations where the approximate methodologies are inadequate, we propose a hybrid method that enables accurate numerical staffing requirements to be produced at a quicker rate.
- We develop integer linear programmes (ILPs) to produce low-cost rosters that satisfy the minimum hourly coverage requirements, which can be solved with practical heuristic algorithms in the workforce planning DSS.
- In the consideration of all the above functions, this research devotes particular attention to the development, solution and validation of sufficiently detailed stochastic models for time-dependent multi-server systems with two customer classes, which can be ultimately employed to optimise resource allocation. Through integrating the steps involved in the rostering process into a single problem, the workforce capacity planning tool developed in conjunction with this research essentially provides a macro view of multiple techniques required to optimise staffing profiles in complex systems with heterogeneous customers and non-stationary demand.

This article is structured as follows: after a brief introduction of WAST and the data provided for analysis is given in Section 2, Section 3 indicates how the forecasting, queueing theory and scheduling/rostering methods developed in the paper are amalgamated in a comprehensive workforce capacity planning tool. Subsections of this section are then used to provide further details of each of the distinct techniques in turn. Specifically, Section 3.1 outlines how accurate demand forecasts, generated using a novel modelling technique known as SSA, are directly fed into the queueing theory models described in Section 3.2. In this section, we also describe how WAST can be approximately modelled as a dual-class time-dependent system and outline how approximate methodologies can be used to generate low-cost staffing profiles by estimating the minimum staffing level needed to exceed specified service levels at various periods throughout the day. Before proposing a hybrid approach which allows exact requirements to be efficiently produced using a combination of approximate and numerical methodologies, we investigate analytical remedies to improve the accuracy of the approximated staffing requirements. Finally, the potential of scheduling and rostering techniques to match personnel resources to fluctuating demand requirements is considered in Section 3.3, before the key contributions of the work are discussed in the conclusion.

2. The research problem

Response times (the interval between arrival of the call and attendance of a paramedic) are one of WAST's Key Performance Indicators (KPIs) since they are believed to provide a good indication of the quality and timeliness of care offered by the service. Maximal acceptable time frames for these are set by Welsh Government and WAST's performance against the targets is reported on a monthly basis.

Distinct targets are specified for different urgencies of emergency requests, which are classified into one of three categories by call takers using a triage system known as the Advanced Medical Priority Dispatch System (AMPDS) (see [2]) as follows:

- Category A—immediately life-threatening condition/injury.
- Category B—serious but not life-threatening condition/injury.
- Category C—neither life-threatening nor serious condition/injury.

The targets, reported by the [19], applied at the time and considered in this research may further be summarised as:

- **Target 1**: To attain and maintain a month on month performance of at least 60% of first responses to Category A calls arriving within 8 min in each Health Board; and to follow up with a fully equipped emergency ambulance to a level of 95% within 14, 18 or 21 min respectively in urban, rural or sparsely populated areas.
- **Target 2**: To send a fully equipped emergency ambulance to all other emergency calls (Category B and Category C) to a level of 95% within 14, 18 or 21 min respectively in urban, rural or sparsely populated areas.

A large fleet of different vehicle types may be called upon by WAST to respond to an emergency request for assistance, but the main vehicles used are Rapid Response Vehicles (RRVs) and fully equipped Emergency Ambulances (EAs). RRVs cannot be used to transport patients as they are typically small vehicles operated by a single health worker; however they offer the advantage that they can rapidly reach the scene of the incident. EAs can be used to transport patients and are typically manned by two crew members (at least one of whom must be a fully trained paramedic). This research assumes that a single EA is sent to **all** emergency calls, and an additional RRV is sent as the first response vehicle to every Category A incident, as should indeed transpire in practice. Accordingly, this paper describes how we develop OR methods to generate recommendations of minimum EA requirements for WAST. The question of the number of RRVs to deploy can be addressed separately, by simplifying some of

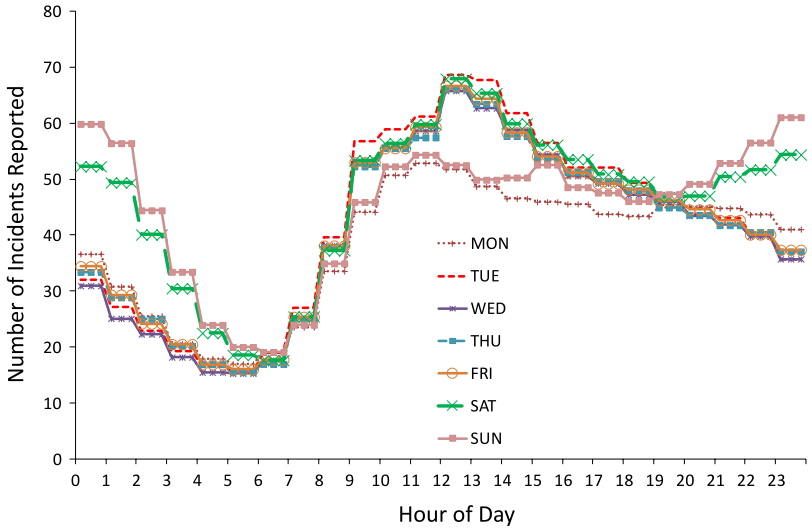

Fig. 1. Mean number of incidents reported per hour, by weekday, Apr 2005–Dec 2009.

the methods described in this paper that deal with two customer classes, to systems that serve a single customer class.

A key consideration that must be taken into account when designing schedules for ambulance crews is that demand for WAST assistance is far from stationary, but heavily dependent upon the time of day and day of week as shown in Fig. 1. Over the 56 month period of data provided by WAST, the number of incidents reported to the service rose from an average of 978 each day in 2005/06 to 1024 in 2008/09, with monthly periodicities, special-day effects, autocorrelations and random fluctuations, as described in [20].

Moreover, since requests for WAST assistance are prioritised according to urgency, all of the techniques that are described to optimise WAST resources in the following sections are accordingly designed to aptly deal with both non-stationary and prioritised demand.

3. Workforce planning

The process of optimising resources by means of rostering of employees using low-cost shifts that match stochastic demand levels requires the investigation of several inter-related procedures. The process traditionally begins with the consideration of methods to generate accurate forecasts of demand, followed by techniques to convert the demand profiles to coverage requirements and generate optimised shift schedules. The resulting shift schedule can be ultimately used as input to a rostering system, detailing the work to be performed over a specified time period by each member of the workforce in a way to minimise labour costs. Most current practice to optimise personnel scheduling follows the general approach originally presented in [21], which recommends that the following steps be taken to roster employees: (i) forecast demand; (ii) convert demand forecasts into staffing requirements; (iii) schedule shifts optimally; and (iv) assign employees to shifts.

The research considered in this article however integrates the processes into a self-contained DSS for WAST, designed to find minimum staffing requirements that allow the government response time targets to be met, as illustrated in Fig. 2. Despite the fact that the tool amalgamates several sophisticated and complex interrelated analytical techniques, it is designed with a user friendly interface in Microsoft Excel. Since Excel software is familiar to planners, WAST employees are able to use the tool to obtain any one, or a combination, of the outputs specified in boxes 1, 3, 4(a) or 4(b) in Fig. 2 (note that box 2 is simply a procedural step that converts the demand forecasts into the staffing requirements

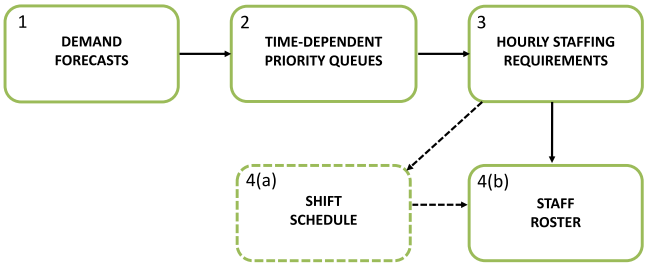

Fig. 2. Integration of techniques in workforce capacity planning tool.

output in box 3). If all options are executed, the end product is a low-cost staff roster that adheres to working time directives and the waiting time targets.

The following sections address each of the procedures presented in Fig. 2 in turn, with a brief description of the methodology developed to deal with the complex pattern of demand for WAST assistance.

3.1. Demand forecasts

Despite the potential of advanced statistical models to offer accurate demand forecasts, WAST currently uses a fairly rudimentary demand pattern analysis technique to predict call volumes, known as average peak demand analysis. The method is intended to provide sufficient capacity to respond to requests for assistance during periods with peak demand levels and operates as follows: for each hourly period of the week (i.e. 168 periods), the number of requests made for an ambulance for the corresponding hour in the 50 weeks preceding it is calculated, before the maximum number in each 10-week period selected to provide 5 'peak' demand values. The average of these 5 values is selected as the average peak demand value, and the number of ambulances deployed for that hour in future weeks is based on the concept that there must be a sufficient number to cope with such demand.

Although the demand pattern analysis technique takes into account some stochastic variation of the inputs, it does not necessarily take into account seasonal variations and other stochastic effects that might arise; and a great deal of information is lost through using summary measures to inform the forecasts in place of the data itself [22]. Thus when developing planning models for EMS systems, researchers have used both regression models to explain the spatial variation of demand and time series models

to account for variations over time. The earliest models, based on multiple regression, were often performed on incomplete data sets with outdated socioeconomic and population data; nevertheless they generated models capable of predicting total yearly demand to a high degree of accuracy (see [23–26]). Since the late 1980s, classical time series models such as Autoregressive Integrated Moving Average (ARIMA) and Holt–Winters (HW) methods, have been used extensively to forecast call volumes (see [27–29]) and specifically applied to ambulance demand in [30].

ARIMA models, originally described in [31], provide a class of models to approximate a time series using a large class of autocorrelation functions after allowing the time series to be stationarised through transformations such as differencing and logging. These models account for temporal dependencies using autoregressive (AR) terms, which are lagged observations of the dependent variable and moving average (MA) terms, which are lagged error terms, as explanatory variables. HW models offer an alternative methodology for generating future predictions of demand that incorporate both trend and seasonal variations, using a set of simple recursions that rely on a weighted average of historical data values, with the more recent values carrying more weight.

Whilst HW and ARIMA methods are successful in overcoming some of the shortfalls of regression techniques such as multicollinearity, autocorrelation and the difficulty of selecting covariates; both are however parametric in nature and require restrictive distributional and structural assumptions, such as stationarity of the data. Empirical studies performed in [32] additionally claim that there is no clear winner amongst the univariate methods to forecast call volumes because they perform differently under various lead times and different workloads. Whilst these traditional methods prove useful for upper-level capacity planning and budgeting, recent advances in location analysis, allowing ambulance deployment strategies to become more flexible and dynamic in nature, call for more responsive predictions of demand and model-free methods to predict call volumes [33,34].

In conjunction with evaluating the potential of conventional time series methods to predict future demand levels, our research responds to the request to produce accurate forecasts whilst adequately accounting for nonstationarities, by analysing the capability of a non-parametric technique known as SSA to account for both trend and seasonality patterns exhibited in the data. SSA generates forecasts using a Singular Value Decomposition (observed to generate high quality forecasts in [35]) and the problems inherent with the traditional methods are not present in SSA as it is able to expose important characteristics of the time-series without requiring either a parametric model, or assumptions concerning the signal or noise [36]. Table 1 illustrates that SSA is capable of producing superior long-term forecasts (especially helpful for EMS planning) and at least comparable short-term forecasts to well-established methods. The table evaluates the quality of rolling forecasts generated for July 2009 by SSA, ARIMA and HW using the Root Mean Square Error (RMSE) and standard deviation (SD; reported in brackets). By decomposing a time series into various elements, and separating the trend and periodic components from structureless noise (i.e. random fluctuations), SSA is able to adequately account for the seasonal and stochastic variations in the data when reconstructing the time series and produce forecasts that simultaneously account for several factors affecting demand. Further details regarding the underpinnings of the SSA technique and its ability to produce forecasts of WAST demand are contained in [20].

In further investigations, SSA has been consistently found to generate accurate forecasts for various months and forecasting horizons; especially for longer-term forecasts which are desired by WAST to set staffing schedules and rosters. For practical purposes, forecasts could additionally be updated as new demand levels are obtained, to allow rostering and scheduling plans to be fine-tuned as their implementation date is approached. Nevertheless, since Table 1 illustrates that SSA is capable of generating particularly high quality long-term predictions, its forecasts are less likely to be subjected to significant revisions or implicate such costly last-minute changes to staffing schedules as other methods.

Table 1
Comparison of model forecasts for daily demand by using RMSE (SD) (July 2009).

Average RMSE	SSA	ARIMA	HW
7-day forecast	44.77 (11.03)	44.69 (13.20)	60.12 (15.87)
14-day forecast	44.25 (4.80)	48.96 (7.84)	63.52 (13.83)
21-day forecast	45.04 (3.31)	50.75 (4.63)	60.87 (12.20)
28-day forecast	45.76 (3.02)	50.74 (3.86)	62.50 (10.73)

In addition to producing high quality forecasts, SSA further benefits from its ability to be easily embedded into a spreadsheet tool. Thus having provided motivation for its use as an accurate tool to predict Welsh ambulance demand; we embed the methodology in the DSS tool (step 1 of Fig. 2). The tool contains options that allow the methodology to be flexibly adjusted to produce forecasts at various levels of granularity, as requested by the user.

3.2. Time-dependent priority queues

With the demand forecasts estimated, the next part of the resource allocation optimisation process involves converting these into minimum staffing requirements, which we address using queueing theory. A considerable body of research has shown that queueing theory can be useful in health care (see [37–39]), but the authors in [10] indicate that there is still a need to develop time-inhomogeneous models that capture the time-dependent arrival patterns of patients. Our work directly responds to this call through modelling WAST as a time-dependent priority queue, where Category A incidents are treated with precedence. Using the demand forecasts output from SSA, we consider approximate and numerical queueing theory techniques to generate minimum coverage requirements that satisfy the response targets.

In our approach, we represent WAST as a time-dependent dual-class priority service system with s servers and an unrestricted waiting line. Category A incidents requesting WAST assistance are treated as High Priority (HP) whilst Category B and C incidents are classed as Low Priority (LP), and processed according to a non-preemptive priority (NPRP) queueing discipline, meaning that ambulance crews may only attend a LP incident if there are no HP emergencies logged in the system awaiting a response. However once a crew has been assigned to attend a LP incident, it cannot be re-routed to attend one of a more serious nature until it completes its service with the current patient. HP customers arrive according to a Poisson process with rate $\lambda_H(t)$ and LP customers arrive with rate $\lambda_L(t)$; so the rate of customers arriving for service at each time t is $\lambda(t) = \lambda_H(t) + \lambda_L(t)$, which are all updated at hourly intervals with the average arrival rate for that interval. Service times are independently and exponentially distributed (not class-dependent) with mean time $\frac{1}{\mu}$. It is assumed that all ambulance crews have identical capabilities, operate under the exhaustive service discipline (meaning that if they are attending to a patient when their shift is scheduled to end, they must first complete their service with that particular patient before leaving), and if multiple crews are available to process a job, each available crew has an equal probability of taking on the job. Our goal is to find a desirable staffing function $s(t)$, which defines the minimum number of EAs (or equivalently, EA crews) that need to be deployed in each hourly interval, to limit the proportion of HP and LP patients waiting longer than targeted 14 min response target time to a maximum

of 5% at all times. This may be expressed via two equations, which must be satisfied at all time points:

$$P(W_{qH}(t) > x_H) \leq 0.05 \quad (1)$$

and

$$P(W_{qL}(t) > x_L) \leq 0.05 \quad (2)$$

where $W_{qH}(t)$ and $W_{qL}(t)$ represent the virtual waiting times of a HP and LP patient arriving at time t respectively and 0.05 denotes the maximum allowed excess wait probability. x_H and x_L are the maximum acceptable waiting times that may pass before an ambulance is mobilised to attend HP and LP incidents (adjusted from 14 min to take into account travel times). Essentially, Eqs. (1)–(2) restrict the likelihood of a HP or LP patient requesting EA assistance at time t having to wait longer than the targeted response times for service, to be no greater than 5%. Our queueing model is based on the assumption that exactly one EA is required to attend each incident reported to WAST which is staffed by two ambulance officers, treated as paired for the purpose of coverage requirements, and referred to as 'crew'.

However, the non-stationary nature of demand for WAST assistance discussed in Section 2 renders the queueing model analytically intractable, i.e. there are no closed-form expressions by which one can evaluate various performance metrics over time. Instead, researchers have developed and compared numerical and approximate approaches to generate staffing profiles for organisations subjected to time-dependent demand. When comparing approximate and numerical methods for time-dependent systems, [40] comments that when the approximate approach is justified, then it should be used because it is simpler and faster; but there are many time-dependent queues, especially in health care, where the approximation approaches will not work well, so other methods are required to provide accurate insights of system behaviour.

In order to overcome the shortfalls of approximation techniques, a tractable numerical approach which allows one to accurately track the probability of an excessive wait for both HP and LP customers in $M(t)/M/s(t)/NPRP$ systems with two customer classes has been recently presented in [41]. Much of the queueing theory proposed within this paper in fact builds upon the preliminary work presented in [41], where Vile et al. describe how the behaviour of time-dependent dual class systems can be evaluated using mixed discrete-continuous time Markov chains, with instantaneous transitions to account for both full staff turnovers and minor adjustments made to the workforce across the course of the day. The authors provide insights into the potential of two methods to match demand and staffing levels: firstly, Euler Pri is an exact numerical method that extends the Euler method derived for a single customer class (see [17,42]) to a priority queue with two customer classes. Secondly, SIPP Pri is an approximation that extends the Stationary Independent Period by Period (SIPP) method discussed in [43] to a priority queue with two customer classes. A comparison of the exact outputs generated against approximate requirements generated by each of the techniques reveals that the staffing levels generated by SIPP Pri are often close to the numerical recommendations, but not identical.

Although SIPP Pri clearly generates estimates of staffing requirements in a far more efficient manner; it is only suitable if results are desired within a reasonable accuracy. Thus for cases where it is essential to obtain accurate staffing requirements, we propose a hybrid approach which allows generation of the exact numerical solution in a shorter computation time. Prior to embedding SIPP Pri in the hybrid methodology, Section 3.2.1 first investigates the potential of three modifications that can be made to the arrival rate function prior to its insertion within SIPP, to improve the accuracy of its predictions.

3.2.1. Variants of SIPP Pri

The SIPP Pri approach outlined in [41] estimates time-dependent behaviour by segmenting the period of operation into distinct shifts, calculating the mean arrival rate in each one, and matching staffing to demand requirements for each shift using a series of stationary closed-form steady-state formulae (assuming that the system operates at a consistent mean level within each shift and is independent of the behaviour in neighbouring periods). In this way, minimum hourly staffing requirements can be obtained by incrementing the number of servers, s, employed for each hourly period until a desired service level is achieved. Since the technique assumes that the staffing levels for each period can be determined independently and steady-state conditions are achieved within each hour, its outputs can be misleading in periods for which these assumptions are violated, as the queue length at the start of each period is in reality heavily reliant upon the number of customers remaining in the system at the end of the previous period.

Previous research works concerning systems with a singular customer class have shown that the accuracy of the standard SIPP approach can be improved by adjusting the arrival rate function prior to its implementation within SIPP. For a comprehensive summary of the effects arising from the application of a wide range of such transforms upon model performance, the reader is referred to [43]. A popular revision to the arrival rate is known as Lag Avg (or Lag SIPP) [44] which uses a modified arrival rate to account for customers who arrived in an earlier period, but receive service in a subsequent period. SIPP Mix has also been proposed to overcome the problem of understaffing when the arrival rate is decreasing, by using the average arrival rates for phases where it is strictly increasing, and the maximum arrival rate otherwise. For the purpose of using SIPP Pri to recommend minimum staffing requirements for WAST and other $M(t)/M/s(t)/NPRP$ systems, we propose three revisions that can be made to the arrival rate prior to its insertion into the SIPP Pri algorithm: the first two are direct extensions of those proposed in [43]; and the third is a more responsive modification we offer to overcome some of the shortfalls of the aforementioned techniques:

- Lag Avg Pri: This directly extends the standard Lag Avg approaches to enable its application within $M(t)/M/s(t)/NPRP$ systems. It estimates the required staffing level based on the average arrival rates predicted for the relevant period shifted back by L units (estimated as the average service time) in an attempt to incorporate the lag that commonly exists between the peak arrivals and peak congestion.
- SIPP Mix Pri: This technique uses the average planning period arrival rates in all periods where the overall arrival rate is strictly increasing, and the maximum arrival rates otherwise (calculated as 1.2 × average rate, based on preliminary investigations), to avoid the problem of understaffing.
- Adaptive SIPP Pri: In recognition that SIPP Pri often recommends staffing levels that are *above* the exact requirements for WAST (see [41]), we propose that a more appropriate arrival rate function could be achieved by taking an average of the rate in the current and preceding period, for periods in which the expected rates differ by more than 20%. This revision aims to incorporate the effect of previous arrivals in each period, whilst avoiding the problem of over/understaffing where the approximation methods fail to recognise that it takes time for the queue to increase/decrease significantly.

Thus the SIPP Pri methodology remains consistent in all the above cases and the only adjustments are those made to the arrival rate function prior to the application of the technique. Table 2 displays the average RMSE associated with the minimum number of EAs recommended to be deployed in South East Wales by each of the approximate methodologies to meet the response time targets outlined in Section 2, when compared to the exact Euler Pri requirements, for each hour of the first four weeks (i.e. 672 hourly

Table 2
Average RMSE of SIPP Pri, Lag Avg Pri, SIPP Mix Pri and Adaptive SIPP Pri staffing recommendations, compared against Euler Pri results (July and Dec 2009).

Hour	$\lambda_H + \lambda_L$ (Avg Rate of Arrival)	SIPP Pri Jul/Dec RMSE	Lag Avg Pri Jul/Dec RMSE	SIPP Mix Pri Jul/Dec RMSE	Adaptive SIPP Pri Jul/Dec RMSE
0	4.8/5.1	0.37/0.62	0.83/0.72	**1.29**/**1.35**	0.53/0.62
1	4.5/4.8	0.42/0.38	0.68/0.80	0.82/0.65	0.63/0.38
2	3.8/4.0	0.33/0.33	0.78/**1.00**	0.89/**1.00**	0.42/0.33
3	3.1/3.3	0.46/0.65	0.96/**1.41**	0.85/**1.25**	0.46/0.76
4	2.2/2.3	0.53/**1.05**	**1.21**/0.73	0.63/0.87	0.53/0.50
5	1.9/2.0	0.53/0.60	0.76/0.57	0.76/0.73	0.65/0.50
6	3.2/3.5	0.76/0.94	**1.13**/**1.13**	0.76/0.94	0.68/0.38
7	3.8/4.2	0.85/0.73	0.38/0.68	0.93/0.82	0.53/0.38
8	5.4/5.9	**1.00**/**1.10**	**1.25**/**1.16**	**1.00**/**1.10**	0.53/0.19
9	6.9/7.4	0.98/**1.05**	0.85/0.78	0.98/**1.05**	0.38/0.19
10	7.7/8.3	0.93/0.82	0.38/0.71	0.93/0.82	0.76/0.63
11	7.5/8.1	0.63/0.65	0.73/**1.27**	**1.68**/**1.74**	0.63/0.65
12	2.9/3.2	**2.43**/**2.34**	**3.07**/**3.16**	**1.77**/**1.52**	**1.20**/**1.07**
13	3.1/3.5	0.65/0.38	0.38/0.27	0.65/0.38	0.65/0.38
14	3.1/3.5	0.00/0.50	0.53/0.19	0.76/0.89	0.00/0.50
15	3.0/3.4	0.38/0.42	0.76/0.46	**1.00**/0.78	0.38/0.42
16	3.1/3.4	0.46/0.42	0.57/0.42	0.96/0.85	0.27/0.19
17	3.0/3.3	0.60/0.19	0.73/0.46	**1.04**/0.80	0.60/0.19
18	3.2/3.5	0.00/0.00	0.65/0.38	0.53/0.53	0.00/0.00
19	4.9/5.3	1.00/0.87	**1.13**/**1.20**	**1.00**/0.87	0.00/0.38
20	4.9/5.3	0.53/0.65	0.53/0.93	0.76/**1.00**	0.53/0.65
21	5.3/5.7	0.57/0.42	0.80/0.42	**1.25**/0.82	0.65/0.19
22	5.1/5.4	0.50/0.38	0.76/0.91	0.98/**1.30**	0.42/0.19
23	5.0/5.3	0.19/0.33	0.57/0.73	0.68/0.73	0.19/0.33
Mean	4.2/4.6	0.78/0.80	**1.00**/**1.03**	**1.00**/**1.00**	0.55/0.48

periods) of July and December 2009. The expressions embedded in the approximate methodologies to evaluate the probability that HP and LP patients wait longer than the acceptable response times are taken from [41]. Whilst closed-form expressions are not readily available to calculate the probability of an excessive wait for two customer classes, all of the approximate techniques still possess the benefit that they are able to estimate the minimum requirements at a rapid rate. For example, whilst Euler Pri requires around 120 min to generate hourly EA requirements for a 3-month forecasting horizon on a 3 GHz machine with 2.96 GB of RAM; SIPP Pri can offer an approximate solution in around 10 min, and the additional time required by the variants of the SIPP Pri technique is only that required to obtain adjusted arrival rate functions.

When each of the variants of the SIPP Pri approach is implemented to generate minimum staffing requirements that comply with the waiting time targets defined in Section 2, for two 28-day test periods (the same as in [41]), Table 2 reveals that our Adaptive SIPP Pri approach generates staffing requirements that are closer to the Euler Pri recommendations than the standard SIPP Pri results (see column 5); but that the Lag Avg Pri and SIPP Mix Pri results (in columns 3–4) are generally inferior. The results indicate that the periods in which SIPP Pri, Lag Avg Pri and SIPP Mix Pri fail to produce reliable requirements for are predominantly 08:00–09:00 and 12:00–13:00. Interpreting these findings in context of the changing arrival rates across the course of the day, as visualised in Fig. 1, it is clear that the main problem in using SIPP Pri to construct requirements for the 08:00–09:00 period is that it fails to recognise the link with previous periods or account for the fact that it takes time for the queue to build up to a level great enough to deploy an additional EA. Contrastingly, the error associated with the 12:00–13:00 period is mainly attributable to underpredictions, since the demand exhibited in this period is considerably lower than the previous period.

In [43], Green et al. have previously found that the non-priority counterpart of Lag Avg Pri performs poorly when the relative amplitude is high; and that standard SIPP Mix commonly understaffs when the arrival rate is decreasing. Since the RA is high for a large portion of the data in our case study, it is unsurprising that our empirical results for Lag Avg Pri results are inferior to the standard SIPP Pri results. Furthermore, Table 2 demonstrates that SIPP Mix Pri only offers marginally improved recommendations since the method is primarily designed to overcome problems associated with overstaffing, whilst understaffing is the most predominant shortfall of SIPP Pri in this investigation. It appears logical that in order to generate accurate results, each of the SIPP Pri extensions requires the same conditions to hold as their non-priority equivalent, i.e. they are reliable if the relative amplitude is low (around 0.1–0.5) and planning periods are short (around 0.25–0.5 h). Since neither of these conditions are strictly met in this case study, it is not surprising that they fail to improve the results of the standard Priority SIPP approach.

Contrastingly, our proposed Adaptive SIPP Pri approach appears highly successful in improving the accuracy of the SIPP Pri predictions. This supports the use of our novel method as a practical technique to improve the approximate approach, and directly demonstrates how improved requirements can be generated in systems where SIPP Pri performs poorly. Due to the methodology followed by the technique, it is equally capable of improving SIPP Pri staffing requirements in systems where overstaffing or understaffing are shortfalls of the standard approach, if the error is attributable to the failure of the technique to recognise the impact of staffing levels and arrival rates in previous periods. In addition to empirically demonstrating that the technique is capable of producing more reliable staffing for WAST in Table 2; we more generally expect the technique to be more robust to higher relative amplitudes than the standard SIPP Pri technique (due to its capacity to adjust the arrival rate in consecutive periods with widely differing arrival rates) and more robust in systems with longer service rates (as it considers the effect of the time-lag that exists between arrival and service times). Further, by accounting for the effect of arrivals in earlier periods in its calculations, we have shown it is able to offer improved predictions if staffing requirements are desired for moderately longer planning periods than 0.25 or 0.5 h, as required by SIPP/SIPP Pri. However the technique should not be used in systems with planning periods spanning several hours, or which exhibit low presented loads, since such systems are more likely to reach steady state within each planning period, so the modifications considered by the technique would be unsuitable to improve the accuracy of predictions.

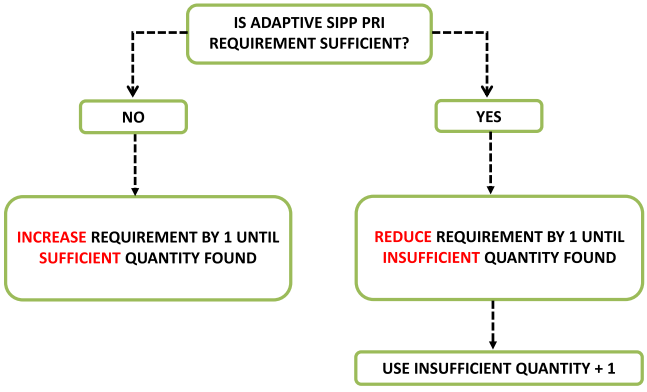

Fig. 3. Line of enquiry followed by hybrid approach to generate staffing requirements.

Table 3
Pool of allowable shifts.

Shift number	Shift time	Shift duration (h)
1	06:00–12:00	6
2	06:00–18:00	12
3	07:00–16:00	9
4	08:00–17:00	9
5	09:00–20:00	11
6	15:00–00:00	9
7	16:00–01:00	9
8	16:00–04:00	12
9	17:00–02:00	9
10	21:00–06:00	9
11	02:00–07:00	5

3.2.2. Hybrid approach

Balancing the ability of the approximation methods to provide rough solutions rapidly and the advantage of the Euler Pri method outlined in [41] to produce accurate predictions at the expense of computation speed, the ultimate methodology we embed in step 2 of DSS for WAST (see Fig. 2) is a hybrid approach which employs both methods to efficiently produce the accurate hourly staffing requirements desired in step 3. If numerical analysis finds that the Adaptive SIPP Pri predictions are sufficient, the Euler Pri method can be used to test if the desired performance level could in fact be achieved with fewer staff, by decrementing the Adaptive SIPP Pri suggested staffing levels for each period in integer steps, until a staffing level is reached that violates the waiting time targets. If the converse is true and the initial quantities are found to be insufficient, the minimum required staffing levels can be obtained by incrementing the initial quantity in integer steps until a sufficient number is found. These steps are summarised in Fig. 3.

Applying this methodology to generate accurate staffing requirements for a free-standing week in July reduces the computation time from 10 to 8 min on a 3 GHz machine with 2.96 GB of RAM. Whilst this is a small saving for this dataset, a 20% reduction in time could become quite considerable if requirements are requested for longer periods.

3.3. Scheduling and rostering

Finally, with the minimum hourly coverage requirements produced, we investigate shift scheduling and rostering techniques that can be used to construct a desirable roster for WAST crews that satisfy the minimum hourly requirements, since, in reality, EA crews cannot be employed for single hourly periods, but only for a combination of a limited number of pre-defined shifts that satisfy Working Time Directives (WTDs). Practically all EMS agencies plan their crew shifts in advance, although shift lengths vary by agency [45]. All teams based in South East Wales currently use 11 predefined shifts that vary between 5 and 12 h in length, as detailed in Table 3.

Section 3.3.1 outlines how we formulate an ILP to select a desirable combination of shifts to ensure that the minimum hourly staffing requirements are covered with the lowest number of labour hours, and Section 3.3.2 details our ILP to assign crews to shifts so as to minimise staff costs. The ILPs are formulated in such a way that enables them to be incorporated as part of the Excel-based capacity planning tool, and solved using heuristic search techniques.

3.3.1. The shift scheduling ILP

Using the minimum number of EAs required for each hourly period generated from the hybrid methodology presented in Section 3.2.2 as input, our shift scheduling ILP seeks to determine a set of appropriate shifts that minimise the total number of labour hours required, under the assumption that each EA is staffed by two ambulance officers, treated as paired and referred to as 'crew' for the purpose of this investigation. The approach can fundamentally be considered as an adaptation of Dantzig's Labour Scheduling Model (DLSM) investigated in [46], which attempts to minimise the total labour cost by allocating shifts subject to the constraint that sufficient EA crews are present in all periods. Defining the sets and variables as:

- D: the set of days in the scheduling horizon
- P: the set of hourly periods in a day
- S: the set of allowable shifts
- x_{sd}: the number of crews working shift s on day d
- r_{sd}: the desired crew requirement in shift s on day d
- r_{pd}: the desired crew requirement in period p on day d
- c_s: the cost of assigning a crew to work shift s
- $a_{sp} = \begin{cases} 1, & \text{if period } p \text{ is included in shift } s \\ 0, & \text{otherwise} \end{cases}$
- l_s: the length (hours) of each shift ($\sum_{p=1}^{24} a_{sp}$) and
- $p_s = \begin{cases} 0.95, & \text{if shift } s \text{ operates for less than 9 h} \\ 1, & \text{if shift } s \text{ operates for exactly 9 h} \\ 1.05, & \text{if shift } s \text{ operates for more than 9 h;} \end{cases}$

our scheduling model can then be written as:
Minimise,

$$Z = \sum_{s\in S}\sum_{d\in D} x_{sd}c_s. \tag{3}$$

Subject to constraints:

$$\sum_{s\in S} x_{sd}a_{sp} \geq r_{pd}, \quad \forall\, p = 1, 2, \ldots, 24,\ d = 1, \ldots, 28 \tag{4}$$

$$x_{sd} \geq 0 \text{ and integer}, \quad \forall\, s = 1, 2, \ldots, 12,\ d = 1, 2, \ldots, 28, \tag{5}$$

$$x_{11,d} = x_{12,d+1}, \quad \forall\, d = 1, 2, \ldots, 27 \tag{6}$$

$$x_{12,1} = 0. \tag{7}$$

The objective function presented in Eq. (3) attempts to minimise the number of crews assigned to each shift by allocating shifts subject to the constraints (4)–(5) so that sufficient employees are present in all periods. Note that since the 10 pm–7 am shift overlaps the day boundary at 6 am, this must be formulated in our model as two separate shifts (namely 10 pm–6 am (the 11th shift input) and 6 am–7 am (the 12th shift input)). Subsequently, the model must specify $s = 1, \ldots, 12$ in place of $s = 1, \ldots, 11$ with the additional constraints (6) and (7) to ensure that any crew assigned to work the last shift on day d (10 pm–6 am) also work the first hour on day $d + 1$ (6 am–7 am).

The weights assigned to the shifts in the objective function can be flexibly adjusted by the planner in the DSS, but the default weights are selected to reflect both the duration and preference of each shift, such that:

$$c_s = l_s \times p_s. \tag{8}$$

Whilst the shift schedule may be optimised prior to the application of a rostering model, our research acknowledges the benefit in simultaneously constructing the shift schedule and roster (i.e. combining step 4(a) with step 4(b) in Fig. 2), due to complex working time directives that can prevent crews from working certain shift patterns of the optimised shift schedule. In Section 3.3.2, we present the formulation of the ILPs we embed in the DSS to construct an optimised roster for a weekly period for the South East region of Wales, and further explain how they can be solved heuristically in Section 3.3.3. In line with the structure of WAST's scheduling procedure, 'days' are considered as running from 6 am to 6 am.

3.3.2. The crew allocation ILP

Rostering ambulance officers is a highly constrained optimisation problem, since workforce planners must take into account a number of legal, managerial and practical requirements when assigning health care workers to shifts [7,47,48]. The imposed requirements can usually be described by two sets of constraints: hard constraints (that must always be satisfied) and soft constraints (which are desirable to be met, but may be violated with a penalisation cost in certain circumstances). The set of constraints differs from Trust to Trust, and those relevant to WAST are discussed in the WTD which are outlined in the 'Agenda for Change' handbook [49].

Our crew allocation ILP model aims to simultaneously reduce the total size of the workforce and total labour hours by considering the assignment of overtime hours (i.e. more than 38 hours per week) as a violation of a soft constraint which is penalised with an additional cost in the objective function. 50 staff are potentially offered to the model for selection, as preliminary investigations show that the quantity needed to satisfy the demand requirements should be far lower than this quantity. In addition to the notation defined above, we define additional sets and variables as:

- J: the set of available ambulance crews ($j = 1, 2, \ldots, 50$)
- overtime_j: the number of overtime hours assigned to each crew j
- $w_{jsd} = \begin{cases} 1, & \text{if crew } j \text{ works shift } s \text{ on day } d \\ 0, & \text{otherwise} \end{cases}$
- $\text{crew}_j = \begin{cases} 1, & \text{if crew } j \text{ is assigned at least one shift in the scheduling horizon} \\ 0, & \text{otherwise.} \end{cases}$

Adhering to the shift coverage requirements, our ILP model aims to minimise staff costs by weighting the total number of staff employed for the scheduling horizon and the number of overtime hours assigned in the objective function, as follows:

Minimise,

$$Y = 25 \sum_{j=1}^{50} \text{crew}_j + \sum_{j=1}^{50} \text{overtime}_j. \tag{9}$$

Subject to constraints:

$$\sum_{j=1}^{50} w_{jsd} = r_{sd}, \quad \forall\, s \in 1, 2, \ldots, 12,\ d \in 1, 2, \ldots, 7 \tag{10}$$

$$\sum_{s=1}^{12} \left(\sum_{d=1}^{7} w_{jsd} \sum_{p=1}^{24} a_{sp} \right) \leq 42, \quad \forall\, j \in 1, 2, \ldots, 50 \tag{11}$$

$$\sum_{s=1}^{12} \left(\sum_{d=1}^{7} w_{jsd} \sum_{p=19}^{24} a_{sp} \right) \leq 8, \quad \forall\, j \in 1, 2, \ldots, 50 \tag{12}$$

$$\sum_{s=1}^{10} w_{j,s,d} \leq 1, \quad \forall\, j \in 1, 2, \ldots, 50,\ d \in 1, 2, \ldots, 7 \tag{13}$$

$$\sum_{s=2}^{11} w_{j,s,d} \leq 1, \quad \forall\, j \in 1, 2, \ldots, 50,\ d \in 1, 2, \ldots, 7 \tag{14}$$

$$\sum_{s=5}^{11} w_{j,s,d} + \sum_{s=1}^{2} w_{j,s,d+1} \leq 1, \quad \forall\, j \in 1, 2, \ldots, 50,\ d \in 1, 2, \ldots, 7 \tag{15}$$

$$\sum_{s=6}^{11} w_{j,s,d} + \sum_{s=1}^{5} w_{j,s,d+1} \leq 1, \quad \forall\, j \in 1, 2, \ldots, 50,\ d \in 1, 2, \ldots, 7 \tag{16}$$

$$\sum_{s=10}^{11} w_{j,s,d} + \sum_{s=1}^{8} w_{j,1,d+1} \leq 1, \quad \forall\, j \in 1, 2, \ldots, 50,\ d \in 1, 2, \ldots, 7 \tag{17}$$

$$w_{j,11,d} + \sum_{s=1}^{9} w_{j,s,d+1} \leq 1, \quad \forall\, j \in 1, 2, \ldots, 50,\ d \in 1, 2, \ldots, 7 \tag{18}$$

$$w_{j,11,d} = w_{j,12,d+1}, \quad \forall\, j \in 1, 2, \ldots, 50,\ d \in 1, 2, \ldots, 7 \tag{19}$$

$$\text{crew}_j \geq w_{jsd} \quad \forall\, j \in 1, 2, \ldots, 50, \quad s \in 1, 2, \ldots, 12,\ d \in 1, 2, \ldots, 7 \tag{20}$$

$$w_{jsd} \in 0, 1 \quad \forall\, j \in 1, 2, \ldots, 50, \quad s \in 1, 2, \ldots, 12,\ d \in 1, 2, \ldots, 7; \tag{21}$$

$$\text{crew}_j \in 0, 1 \quad \forall\, j \in 1, 2, \ldots, 50. \tag{22}$$

The coefficient of crew_j in the objective function (9) may of course be flexibly adjusted to appropriately preference the goal of reducing the size of the workforce over the number of overtime hours assigned, but the default coefficient assigned is built upon the assumption that overtime hours are paid at a rate 1.5 times higher than the standard wage; so as the standard working week consists of 38 h, the weekly pay for 1 full member of staff equates to 25 overtime hours.

Constraint (10) ensures that sufficient crews are assigned to each shift to satisfy the coverage requirements and constraints (11)–(19) ensure that strict WTDs are adhered to. In particular, constraint (11) ensures that the maximum number of hours worked in the 7-day period does not exceed 42 h and constraint (12) prevents crews from working more than 8 night hours per week (classed as from midnight to 5 am inclusive i.e. periods 19–24 for a day considered to operate from 6 am to 6 am). Constraints (13)–(18) ensure that all crews receive at least 11 h rest break between shifts, by preventing them from working more than 1 shift per 'day' or being allocated specific tours that violate this WTD. Constraint (19) ensures that the same crew is assigned to the 11th shift on day d and 12th shift on day $d + 1$ (essentially the same shift). Finally Eq. (20) defines the dummy variable constructed to count the number of staff employed for at least one shift over the 7-day period and Eqs. (21)–(22) specify w_{jsd} and crew_j as binary variables.

3.3.3. Solving the ILP heuristically

The computational difficulty intrinsic to highly constrained health care scheduling problems has encouraged the development of heuristic approaches [50,51,9]. Although small instances of our scheduling and rostering ILPs can be solved optimally, we correspondingly also propose a heuristic that can be embedded within the Excel-based DSS to offer good quality, though not necessarily optimal, solutions. The specific heuristic we propose is a Simulated Annealing algorithm which aims to find a good quality solution in reasonable computation time by allowing multiple swaps within iterations—whether this be between the shifts

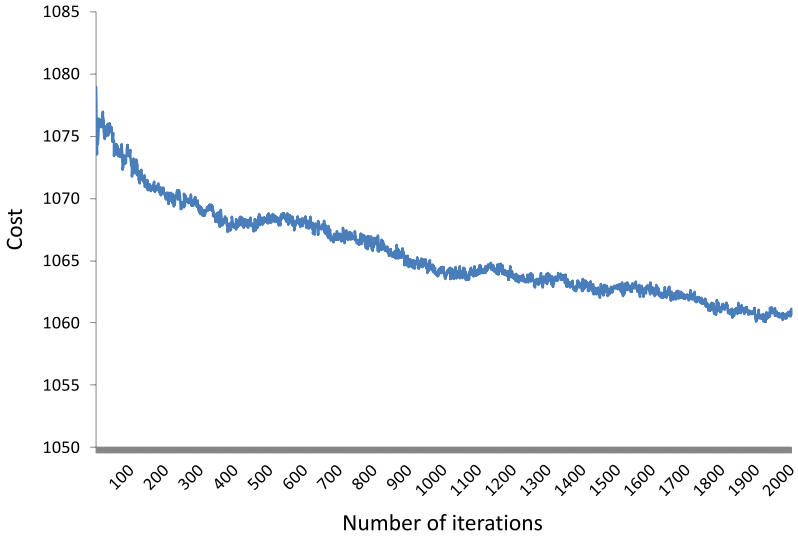

Fig. 4. Rate at which SA heuristic converges (averaged over 50 trials).

themselves or the assignment of shifts to employees to minimise the cost in the weighted objective function defined in Eq. (9).

Fig. 4 shows the average cost output by the SA algorithm (over 50 trials) as it converges to reach a minimum cost for the objective function, given constraints (10)–(22) and hourly staffing requirements for a 1-week period.

The chart shows that the optimal cost selected by the model after 2000 iterations is 1061, although average *best* solution found for each run is actually lower than this (equal to 1054); since the best solution is not necessarily the cost selected at the termination point of the SA algorithm. This average optimal cost value of 1054 is within 5% of the true optimal solution found using XPress-MP software. However, the real benefit of the heuristic lies in its ability to produce good quality rosters for large problem instances with complex objective functions and multiple constraints, which ILP solvers may lack sufficient power to solve.

3.4. The workforce capacity planning tool

Each of the methodologies that have been described in Sections 3.1–3.3 are ultimately amalgamated together in an Excel-based workforce capacity planning and scheduling tool, with options to execute the demand forecasting, shift scheduling or rostering algorithms individually or sequentially. The tool is purposefully designed with a user-friendly interface with parameters that may be flexibly adjusted by the user to provide staffing recommendations for various scenarios that satisfy the response time targets e.g. the user is able to change/add shifts to the potential pool and adjust parameter values for constraints such as maximum weekly working hours, which are automatically read into the algorithms when executed. While taking into account the importance of accurately estimating future demand, the need to develop OR methodology to evaluate service quality in time-dependent priority multi-server systems, and generate efficient shift schedules, the tool:

A. Incorporates time-series methods that adequately account for the stochastic nature of demand to produce accurate forecasts of future demand;
B. Provides both accurate and approximate evaluations of system performance over time;
C. Permits a certain service quality to be met as inexpensively as possible by generating an efficient staffing function that accurately matches resources to fluctuating demand levels;
D. Assigns staff to shifts in an efficient manner, whilst adhering to governmental regulations and working time directives;
E. Is user-friendly and practical; so it could be used to inform WAST staffing decisions and readily adopted by planners to optimise resources independently.

4. Illustrative example

Table 4 presents an example of a roster generated by the workforce capacity tool for crews in South East Wales for a free-standing week in July, based on reducing the cost in objective function (9), subject to constraints (10)–(22) and the set of allowable shifts detailed in Table 3. In this example, a feasible schedule requiring 39 crews, who are each required to work an average of 38.56 h, is found to be sufficient to adequately cover the forecasted demand. Since the rostering ILP is solved heuristically, marginally different timetables are inevitably output each time the SA algorithm is executed, but the average cost achieved in 50 trials for this problem instance was only 0.5% higher than the true optimal found using XPress-MP software. Whereas ILP solvers may however lack sufficient power to find optimal solutions for large problem instances with complex constraints and objective functions, our proposed SA heuristic may be flexibly adjusted to generate low-cost rosters for any problem instance.

Furthermore, the workforce capacity scheduling tool is programmed in a way that allows optimised staffing profiles to be generated for up to 8000 hourly periods (around 4 months of data, depending on the calculation interval chosen for the Euler Pri methodology). The only restrictions are that the working day is considered as running from 6 am to 6 am, and due to the nature of the staffing constraints, days must be considered in their entirety (i.e. all 24 periods must be accounted for on every day entered into the model). It is however worth noting that forecasting horizons spanning several months considerably increase the time required for its execution. As an indication of the rough timings required to execute each of the staffing programmes using the default parameter values programmed in the model for the 1-week scheduling horizon above, in addition to those required for a 3-month scheduling horizon, Table 5 contains a summary of the approximate times required for the various functions when run on 3 GHz machine with 2.96 GB of RAM:

Thus as the scheduling horizon is increased, the run times required to execute each of the programmes considerably lengthen. Whilst the accuracy of the period requirements output for longer

Table 4
An example weekly schedule for Cardiff EA staff, July 2009.

Crew	Shifts assigned to each crew for each day of July						
	1st	2nd	3rd	4th	5th	6th	7th
1	9	–	3	4	11	–	9
2	1, 11	–	11	–	4	5	–
3	4	1	11	–	4	–	8
4	–	1	5	4	1, 11	11	–
5	–	3	8	–	6	8	–
6	2	–	1, 11	11	–	–	3
7	–	1	1, 11	11	–	1	4
8	1	–	4	10	9	–	4
9	11	11	–	4	3	–	4
10	9	–	1	11	–	4	9
11	1	–	9	10	–	4	6
12	–	8	–	4	9	9	–
13	1	4	–	1	–	10	9
14	11	–	4	9	–	–	9
15	6	–	10	9	–	1	6
16	–	11	–	3	1	3	1, 11
17	–	4	–	4	4	1, 11	11
18	–	9	9	–	9	–	3
19	–	8	9	–	4	9	–
20	4	11	–	8	–	3	1
21	–	9	–	1	9	9	9
22	3	9	–	9	8	–	–
23	11	–	1	1, 11	–	3	4
24	4	–	9	–	1	–	11
25	8	11	–	–	1	6	–
26	1, 11	–	1, 11	–	1	6	–
27	1	–	4	–	10	9	–
28	9	11	–	1	3	9	–
29	–	1	4	1, 11	–	1	1, 11
30	9	9	9	9	–	–	1
31	–	1	10	–	1	4	9
32	8	6	–	8	–	–	1
33	9	–	8	–	9	–	3
34	4	3	1	9	–	–	10
35	9	9	–	1	11	–	–
36	4	–	8	8	–	1	–
37	–	9	9	9	9	–	1
38	–	4	4	9	9	11	–
39	4	3	–	4	11	11	–

Table 5
Run times required to execute programmes for various forecasting horizons.

Programme	Forecasting horizon	
	1 week	3 months
Generate SSA demand forecast	3 min	5 min
Compute period requirements (SIPP Pri)	0.3 min	10 min
Compute period requirements (Euler Pri)	10 min	120 min
Compute period requirements (Hybrid)	8 min	100 min
Produce optimised shift schedule	0.5 min	7 min
Produce optimised roster	50 min	180 min

scheduling horizons should not be compromised, the quality of the shift schedule and roster is potentially poorer unless the default parameter values programmed in the tool for the shift scheduling and rostering algorithms are adjusted accordingly. For example, the average cost achieved in 50 executions of the SA algorithm using the default parameter values for the 1-week period described above was only 0.5% higher than the true optimal, whereas the average cost achieved for schedules developed for 3-month periods was around 3% higher.

5. Conclusions

This paper illustrates the ways in which OR can assist with EMS planning. Using a range of modelling tools, we have described our interactions with WAST and outlined several methodologies to aid planners with decisions surrounding the optimal deployment of resources. Being keen to develop new initiatives to improve performance, senior managers at WAST have offered directions for research throughout the project and have commended the ultimate set of highly user friendly tools developed to support future capacity planning decisions. The Clinical R&D Manager at WAST recently reinforced the usefulness of the complex mathematical modelling investigations, commenting "The work is an extremely relevant contribution to implementing policy and procedural changes at WAST", and the Welsh Government have further stated their wish to oversee the implementation of the developed tools to support WAST going forward. If successful, the generic nature of the modelling techniques considered means the tool could further be adopted by ambulance services internationally and used to improve the quality of care provided to patients.

Moreover, because the workforce planning tool has been programmed in a generic fashion with a user friendly interface, there are opportunities to apply the tool to services beyond the EMS. The methodology could in fact be applied to any service concerned with determining minimum staffing requirements that limit the proportion of customers waiting longer than targeted response times to predefined thresholds, such as call centres or A&E departments to name but a few, and it would be interesting to additionally investigate the potential of the tool to improve resource allocation within these organisations in future work.

From an OR perspective, the unique linking together of the techniques in a planning tool which further captures time-dependency and two priority classes enables this research to outperform previous approaches, which have generally only considered a single class of customer, or generated staffing recommendations using approximation methods that are only reliable under limited conditions. In particular, the research has proposed a hybrid approach which enables accurate minimum staffing recommendations to be efficiently generated for $M(t)/M/s(t)/NPRP$ systems, which are widespread throughout industry and commonly expected to attain minimum performance standards. In future work, the practical contribution offered by master capacity planning tool could be further extended by including real-time online analysis of data to allow for short term adjustments to staffing recommendations in response to unforeseen external factors which may influence the demand and staffing profiles.

Acknowledgments

This research was funded by EPSRC grant EP/F033338/1 (part of the LANCS initiative) and the data underpinning the project was provided by the Welsh Ambulance Services NHS Trust (WAST). The authors would particularly like to thank WAST Research Development Manager Richard Whitfield for his profound support throughout the project and the EURO Summer Institute XXXI conference organisers for providing the opportunity to improve the paper in such a stimulating research environment.

References

[1] Welsh Ambulance Services NHS Trust, Annual report 2011/2012: Improving care. Improving lives, Tech. Rep., 2012.
[2] Lightfoot Solutions, Time to make a difference: Transforming ambulance services in Wales. A modernisation plan for ambulance services and NHS Direct Wales, Final Report, Tech. rep.,(2007). URL: http://www.ambulance.wales.nhs.uk/assets/documents/c4cc0416-9fab-4dea-8753-247a9431c4c763344635912373375O.pdf (accessed 02.06.10).
[3] Welsh Government, Ambulance Services in Wales: June 2012, Tech. Rep. SDR 113/2012, 2012.
[4] Welsh Government, Ambulance Services in Wales: September 2012, Tech. Rep. SDR 187/2012, 2012.
[5] V. Knight, P. Harper, L. Smith, Ambulance allocation for maximal survival with heterogeneous outcome measures, OMEGA—Int. J. Manag. Sci. 40 (6) (2012) 918–926.
[6] V. Knight, P. Harper, Modelling emergency medical services with phase-type distributions, Health Syst. 1 (2012) 58–68.

[7] R. Bruni, P. Detti, A flexible discrete optimization approach to the physician scheduling problem, Oper. Res. Health Care 3 (4) (2014) 191–199.
[8] A. Ernst, H. Jiang, M. Krishnamoorthy, B. Owens, D. Sier, An annotated bibliography of personnel scheduling and rostering, Ann. Oper. Res. 127 (2004) 21–144.
[9] J. Van den Bergh, J. Belien, P. De Bruecker, E. Demeulemeester, L. De Boeck, Personnel scheduling: A literature review, European J. Oper. Res. 226 (3) (2013) 367–385.
[10] C. Lakshmi, S.A. Iyer, Application of queueing theory in health care: A literature review, Oper. Res. Health Care 2 (1–2) (2013) 25–39.
[11] Z. Askin, M. Armony, V. Mehrota, The modern call center: A multi-disciplinary perspective on operations management research, Prod. Oper. Manage. 16 (6) (2007) 665–688.
[12] N. Gans, G. Koole, A. Mandelbaum, Telephone call centers: Tutorial, review and research prospects, Manuf. Serv. Oper. Manage. 5 (2003) 79–141.
[13] D. Matteson, M. McLean, D. Woodard, S. Henderson, Forecasting emergency medical service call arrival rates, Ann. Appl. Stat. 5 (2B) (2011) 1379–1406.
[14] H. Setzler, S. Park, C. Saydam, EMS call volume predictions: A comparative study, Comput. Oper. Res. 36 (2009) 1843–1851.
[15] L. Green, P. Kolesar, W. Whitt, Coping with time-varying demand when setting staffing requirements for a service system, Prod. Oper. Manage. 16 (2007) 13–39.
[16] M. Defraeye, I. Van Nieuwenhuyse, Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm, Decis. Support Syst. 54 (4) (2013) 1558–1567.
[17] N. Izady, D. Worthington, Setting staffing requirements for time dependent queueing networks: The case of Accident and Emergency departments, European J. Oper. Res. 219 (3) (2012) 531–540.
[18] B. Chen, S. Henderson, Two issues in setting call centre staffing levels, Ann. Oper. Res. 108 (2001) 175–192.
[19] Welsh Government, Ambulance Services in Wales: February 2011, Tech. Rep. SDR 59/2011, 2011.
[20] J. Vile, J. Gillard, P. Harper, V. Knight, Predicting ambulance demand using singular spectrum analysis, J. Oper. Res. Soc. 63 (11) (2012) 1556–1565.
[21] E. Buffa, M. Cosgrove, B. Luce, An integrated work shift scheduling system, Decis. Sci. 7 (1976) 620–630.
[22] J. Gillard, V. Knight, Using singular spectrum analysis to obtain staffing level requirements in emergency units, JORS 65 (5) (2014) 735–746.
[23] C. Aldrich, J. Hisserich, L. Lave, An analysis of the demand for emergency ambulance service in an urban area, Am. J. Public Health 61 (1971) 1156–1169.
[24] K. Siler, Predicting demand for publicly dispatched ambulances in a metropolitan area, Health Serv. Res. 10 (3) (1975) 254–263.
[25] T. Kvalseth, J. Deems, Statistical models of the demand for Emergency Medical Services in an urban area, Am. J. Public Health 69 (3) (1979) 250–255.
[26] R. Kamentzky, L. Shuman, H. Wolfe, Estimating need and demand for prehospital care, Oper. Res. 30 (6) (1982) 1148–1167.
[27] L. Bianci, J. Jarrett, C. Hanumara, Forecasting incoming calls to telemarketing centers, J. Bus. Forecast. 12 (2) (1993) 3–11.
[28] B. Andrews, S. Cunningham, L.L. Bean, improves call-center forecasting, Interfaces 25 (6) (1995) 1–13.
[29] J. Holcomb, N. Sharpe, Forecasting police calls during peak times for the city of Cleveland, CS-BIGS 1 (1) (2007) 47–53.
[30] N. Channouf, P. L'Ecuyer, A. Ingolfsson, A. Avramidis, The application of forecasting techniques to modelling Emergency Medical System calls in Calgary, Alberta, Health Care Manag. Sci. 10 (1) (2007).
[31] G. Box, G. Jenkins, Time Series Analysis, Forecasting and Control, Holden-Day, Incorporated, San Francisco, 1970.
[32] J. Taylor, A comparison of univariate time series methods for forecasting intraday arrivals at a call center, Manage. Sci. 54 (2008) 253–265.
[33] L. Brotcorne, G. Laporte, F. Semet, Ambulance location and relocation models, European J. Oper. Res. 147 (2003) 451–463.
[34] H. Smith, G. Laporte, P. Harper, Locational analysis: highlights of growth to maturity, J. Oper. Res. Soc. 60 (2009) S140–S148.
[35] H. Shen, J. Huang, Interday forecasting and intraday updating of call center arrivals, Manuf. Serv. Oper. Manage. 10 (3) (2008) 391–410.
[36] N. Golyandina, V. Nekrutkin, A. Zhigljavsky, Analysis of Time Series Structure: SSA and Related Techniques, Chapman & Hall/CRC, New York, London, 2001.
[37] R. Nosek, J. Wilson, Queueing theory and customer satisfaction: a review of terminology, trends, and applications to pharmacy practice, Hosp. Pharm. 36 (2001) 275–279.
[38] J. Preater, Queues in health, Health Care Manag. Sci. 5 (2002) 283.
[39] S. Fomundam, J. Herrmann, A survey of queueing theory applications in health care, ISR Technical Report 24.
[40] A. Ingolfsson, F. Campello, X. Wu, E. Cabral, Combining integer programming and the randomisation method to schedule employees, European J. Oper. Res. 202 (1) (2010) 153–163.
[41] J. Vile, J. Gillard, P. Harper, V. Knight, A queueing theoretic approach to limit excessive waiting times in time-dependent dual-class service systems, 2015, submitted for publication. Text available at: http://www.julievile.co.uk/publications.html.
[42] A. Ingolfsson, E. Akhmetshina, S. Budge, Y. Li, X. Wu, A survey and experimental comparison of service-level-approximation methods for nonstationary $M(t)/M/s(t)$ queueing systems with exhaustive discipline, INFORMS J. Comput. 19 (2) (2007) 201–214.
[43] L. Green, P. Kolesar, J. Soares, Improving the SIPP approach for staffing service systems that have cyclic demands, Oper. Res. 49 (2001) 549–564.
[44] L. Green, P. Kolesar, The pointwise stationary approximation for queues with nonstationary arrivals, Manage. Sci. 37 (1991) 84–97.
[45] H. Rajagopalan, C. Saydam, E. Sharer, H. Setzler, Ambulance deployment and shift scheduling: An integrated approach, J. Serv. Sci. Manage. 4 (2011) 66–78.
[46] G. Thompson, Labor staffing and scheduling models for controlling service levels, Nav. Res. Logist. 44 (8) (1997) 719–740.
[47] E. Burke, P. De Causmaecker, G. Berghe, A. Van Landeghem, The state of the art of nurse rostering, J. Sched. 7 (2004) 441–499.
[48] S. Petrovik, G. Berghe, A comparison of two approaches to nurse rostering problems, Ann. Oper. Res. 194 (1) (2012) 365–384.
[49] The NHS Staff Council, NHS Terms and Conditions of Service Handbook, Amendment Number 24, Pay Circular (Agenda for Change), March 2011.
[50] I. Marquesa, M. Captivo, M. Pato, Scheduling elective surgeries in a Portuguese hospital using a genetic heuristic, Oper. Res. Health Care 3 (2014) 59–72.
[51] Y. Liu, C. Chu, K. Wang, A new heuristic algorithm for the operating room scheduling problem, Comput. Ind. Eng. 63 (3) (2011) 865–871.